Section 3.2

Check Your Understanding, page 168:

1. The slope is 40. For each additional week, we predict that a rat will gain 40 grams of weight.
2. The *y*-intercept is 100. The predicted weight for a newborn rat is 100 grams.
3. After 16 weeks, we predict the rat's weight to be $\hat{y} = 100 + 40(16) = 740$ grams.
4. The time is measured in weeks for this equation, so 2 years becomes 104 weeks. We then predict the rat's weight to be $\hat{y} = 100 + 40(104) = 4260$ grams, which is equivalent to 9.4 pounds (about the weight of a large newborn human). This is unreasonable and is the result of extrapolation.

Check Your Understanding, page 172:

The answer is given in the text.

Check Your Understanding, page 174:

1. The predicted price for this truck is $\hat{y}$ = 38,257 – 0.1629(8359) = 36,895. The residual is $y - \hat{y}$ = 31,891 – 36,895 = –$5004.
2. The actual price of this truck is $5004 less than predicted based on the number of miles it has been driven.
3. The line over predicts the price the most for the truck with 44,447 miles and a price of $22,896. This truck has a residual of –$8120, which means that the line over predicted the price by $8120. No other truck had a residual that was farther below 0 than this one.

Check Your Understanding, page 176:

1. The backpack for this hiker was almost 4 pounds heavier than expected based on the weight of the hiker.
2. Because there appears to be a negative-positive-negative pattern in the residual plot, a linear model is not appropriate for these data.

Exercises, page 193:

3.35 The equation is $\hat{y} = 80 - 6x$ where $\hat{y}$ = the estimated weight of the soap and x = the number of days since the bar was new.

3.36 The professor believes that the slope is 1 and predicts that an IQ score of 100 will result in a reading score of 50. This means that 50 = a + 1(100). Solving for a, we get a = –50. Thus, the equation is $\hat{y} = -50 + x$ where $\hat{y}$ = the predicted reading test score and x = a child's IQ.

3.37 (a) The slope is 1.109. For each 1 mpg increase in city mileage, the predicted highway mileage will increase by 1.109 mpg.
(b) The *y* intercept is 4.62 mpg. This value is not statistically meaningful because this would represent the highway mileage for a car that gets 0 mpg in the city. There are no cars that get such poor gas mileage.
(c) With city mpg of 16, the predicted highway mpg is $4.62 + 1.109(16) = 22.36$ mpg.

3.38 (a) The slope is 0.882. For each one-point increase in IQ, the predicted reading score will increase by 0.882.

(b) The y intercept is –33.4. This value is not statistically meaningful because this would represent the predicted reading score for a child with an IQ of 0. There are no children with IQs this low.
(c) The predicted reading score for a child with an IQ score of 90 is $-33.4 + 0.882(90) = 45.98$.

3.39 (a) The slope is −0.0053. For each additional week in the study, the predicted pH decreased by 0.0053 units. Thus, the acidity of the precipitation increased over time.
(b) The y intercept is 5.43. The predicted pH level at the beginning of the study (weeks = 0) is 5.43.
(c) At the end of the study, pH is predicted to be $5.43 - 0.0053(150) = 4.635$.

3.40 (a) The slope is –19.87. For each additional 1 degree increase in the average monthly temperature, the predicted amount of gas consumed in Joan's home decreases by 19.87 cubic feet.
(b) The y intercept is 1425. When the average monthly temperature is 0°F, the predicted gas consumption for Joan's home is 1425 cubic feet. This prediction is an extrapolation because the data only included months with an average temperature of more than 20°F. We can't be sure that our linear model will apply to months with temperatures this cold.
(c) $\partial\text{gas} = 1425 - 19.87(30) = 828.9$ cubic feet. We predict that the amount of natural gas Joan will use in a month with an average temperature of 30°F is 828.9 cubic feet.

3.41 No. This would be an extrapolation because the data was collected weekly for only 150 weeks. 1000 months corresponds to about 4000 weeks, which is well outside the observed time period. We can't be sure that the linear relationship continues after 150 weeks.

3.42 No. This would be an extrapolation because the average temperatures for the months in which data were collected varied from about 27°F to 57°F. 65°F is outside of this interval of temperatures. We can't be sure that the linear relationship continues above 57°F.

3.43 The tables below show the predicted values, residuals, and squared residuals for each proposed model.

$\hat{y} = 1 - x$

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
–1	2	2	0	0
1	0	0	0	0
1	1	0	1	1
3	–1	–2	1	1
5	–5	–4	–1	1
				sum = 3

$\hat{y} = 3 - 2x$

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
–1	2	5	3	9
1	0	1	–1	1
1	1	1	0	0
3	–1	–3	2	4
5	–5	–7	2	4
				sum = 18

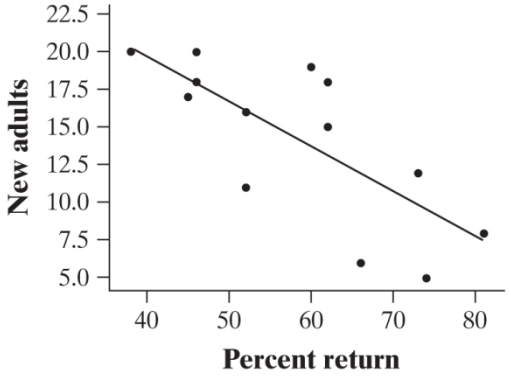
The line $\hat{y} = 1 - x$ is a much better fit. The sum of squared residuals for this line is only 3 while the sum of squared residuals for $\hat{y} = 3 - 2x$ is 18.

3.44 There are many lines we could use to predict gas consumption from temperature. The least-squares line is the line that makes the sum of the squared prediction errors as small as possible.

3.45 The predicted value for this week is $\hat{y} = 5.43 - 0.0053(50) = 5.165$. The residual is $y - \hat{y} = 5.08 - 5.165 = -0.085$. This means that the actual pH value for that week was 0.085 less than predicted.

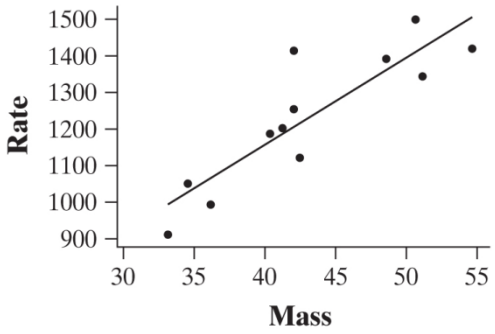
3.46 The predicted value for this point is $\hat{y} = 1425 - 19.87(46.4) = 503.032$. The residual is $490 - 503.032 = -13.032$. This means that the actual amount of gas consumed was 13.032 cubic feet less than predicted in the month of March.

3.47 (a) The scatterplot (with regression line) is shown below.



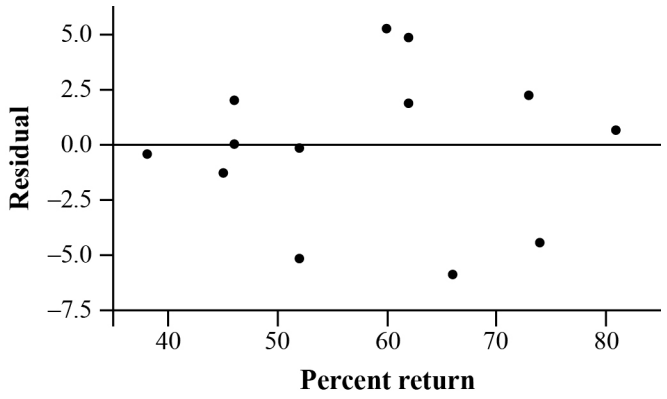
(b) The equation of the least-squares regression line is $\hat{y} = 31.9 - 0.304x$.
(c) For each increase of 1 in the percent of returning birds, the predicted number of new adult birds will decrease by 0.304.
(d) The predicted value for $x = 52$ is $\hat{y} = 31.9 - 0.304(52) = 16.092$. The residual is $y - \hat{y} = 11 - 16.092 = -5.092$. In this colony, there were 5.092 fewer new adults than expected based on the percent of returning birds.

3.48 (a) The scatterplot (with regression line) is shown below.



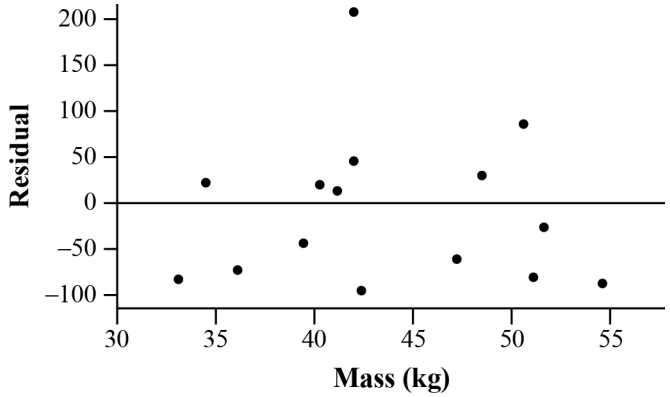
(b) The least-squares regression equation is $\hat{y} = 201.2 + 24.026x$.
(c) For each additional kilogram of body mass, the predicted metabolic rate increases by about 24 cal/day.
(d) The predicted value for $x = 50.6$ is $\hat{y} = 201.2 + 24.026(50.6) = 1416.9$. The residual is $y - \hat{y} = 1502 - 1416.9 = 85.1$. This woman's actual metabolic rate was 85.1 cal/day higher than predicted based on her lean body mass.

3.49 (a) Because there is no obvious leftover pattern in the residual plot, a line is an appropriate model to use for these data.



(b) The point with the largest residual (66% returning) has a residual of about –6. This means that the colony with 66% returning birds has about 6 fewer new adults than predicted based on the percent returning.

3.50 (a) Because there is no obvious leftover pattern in the residual plot, a line is an appropriate model to use for these data. There is one large, positive, outlier, but since it is near the mean of the mass values, it does not influence the line very much.



(b) The point with the largest residual (mass = 42) has a residual of about 200. This means that the person with a lean body mass of 42 kg has a metabolic rate that is about 200 cal/day higher than predicted based on the person's lean body mass.

3.51 No. Because there is an obvious negative-positive-negative pattern in the residual plot, a linear model is not appropriate for these data. A curved model would be better.

3.52 No. Because there is an obvious positive-negative-positive pattern in the residual plot, a linear model is not appropriate for these data. A curved model would be better.

3.53 (a) There is a positive, linear association between the two variables. There is more variation in the field measurements for larger laboratory measurements. Also, the values are scattered above and below the line $y = x$ for small and moderate depths, indicating strong agreement, but the field measurements tend to be smaller than the laboratory measurements for large depths.

(b) No. The points for the larger depths fall systematically below the line $y = x$ showing that the field measurements are too small compared to the laboratory measurements.
(c) In order to minimize the sum of the squared vertical distances from the points to the regression line, the top right part of the line in the scatterplot would need to be pulled down to go through the "middle" of the group of points that are currently below the $y = x$ line. Thus, the slope would decrease and the y intercept would increase.

3.54 Because there is no obvious curved pattern leftover in the residual plot, a linear model is appropriate for these data. However, predictions using the line will be less accurate for larger measurements than for smaller measurements. This is because the residuals are more spread out from the residual = 0 line for larger measurements.

3.55 (a) The predicted free skate score is $\hat{y} = -16.2 + 2.07(78.5) = 146.295$. The residual is $y - \hat{y} = 150.06 - 146.295 = 3.765$. Yu-Na Kim's free skate score was 3.765 points higher than predicted based on her short program score.
(b) Because there is no leftover pattern in the residual plot, a linear model is appropriate for these data.
(c) When using the least-squares regression line with x = short program score to predict y = free skate score, we will typically be off by about 10.2 points.
(d) About 73.6% of the variation in free skate scores is accounted for by the linear model relating free skate scores to short program scores.

3.56 (a) The predicted height is $\hat{y} = 106.1 + 4.21(10) = 148.2$. The residual is $y - \hat{y} = 141 - 148.2 = -7.2$. This student's height was 7.2 cm less than predicted based on the student's age.
(b) Because there is no leftover pattern in the residual plot, a linear model is appropriate for these data.
(c) When using the least-squares regression line with x = age to predict y = height, we will typically be off by about 8.61 cm.
(d) About 27.4% of the variation in height is accounted for by the linear model relating height to age.

3.57 r^2: About 56% of the variation in the number of new adults is accounted for by the linear model relating number of new adults to the percent returning. s: When using the least-squares regression line with x = percent returning to predict y = number of new adults, we will typically be off by 3.67 adults.

3.58 r^2: About 76.8% of the variation in the metabolic rate is accounted for by the linear model relating metabolic rate to lean body mass. s: When using the least-squares regression line with x = lean body mass to predict y = metabolic rate, we will typically be off by 95.08 cal/day.

3.59 (a) The regression line is $\hat{y} = 266.07 - 6.650x$, where y = percent of males that return the next year and x = number of breeding pairs. Following a season with 30 breeding pairs, we find $\hat{y} = 266.07 - 6.650(30) = 66.57$, so we predict that about 67% of males will return.
(b) This is given in the Minitab output as R-Sq = 74.6%
(c) Knowing that $r^2 = 74.6\%$, we find $r = -\sqrt{0.746} = -0.864$. The sign is negative because these variables have a negative association, as indicated by the negative slope.
(d) When using the least-squares regression line with x = number of breeding pairs to predict y = percent returning, we will typically be off by 7.76%.

3.60 (a) The regression equation is $\hat{y} = -0.126 + 0.0608x$, where y = brain activity and x = social distress score. For a person with a social distress score of 2.0, we find $\hat{y} = -0.126 + 0.0608(2) = -0.0044$. The predicted variation in brain activity for this person is –0.0044.
(b) This is given in the Minitab output as $\text{R-sq} = 77.1\%$.
(c) Knowing that $r^2 = 0.771$, we find $r = +\sqrt{r^2} = 0.88$. The sign is positive because these variables have a positive association, as indicated by the positive slope.
(d) When using the least-squares regression line with x = social distress score to predict y = brain activity, we will typically be off by 0.0251.

3.61 (a) The slope is $b = 0.5\left(\frac{2.7}{2.5}\right) = 0.54$. The y intercept is $a = 68.5 - 0.54(64.5) = 33.67$. So the equation for predicting y = husband's height from x = wife's height is $\hat{y} = 33.67 + 0.54x$.
(b) If the value of x is one standard deviation below $\bar{x}$, the predicted value of y will be r standard deviations of y below $\bar{y}$. So, the predicted value for the husband is 68.5 – 0.5(2.7) = 67.15 inches.

3.62 (a) The slope is $b = 0.596\left(\frac{15.35}{5.36}\right) = 1.707$. The y intercept is $a = 9.07 - 1.707(1.75) = 6.083$. So the equation for predicting y = percent change in the index for the entire year from x = percent change in the index in January is $\hat{y} = 6.083 + 1.707x$.
(b) If the value of x is two standard deviations above $\bar{x}$, the predicted value of y will be $2r$ standard deviations of y above $\bar{y}$. So, the predicted value for the percent change for the entire year is 9.07 + 2(0.596)(15.35) = 27.4%.

3.63 (a) $r^2 = (0.5)^2 = 0.25$. About 25% of the variation in husbands' heights is accounted for by the linear model relating husband's height to wife's height.
(b) When using the least-squares regression line with x = wife's height to predict y = husband's height, we will typically be off by 1.2 inches.

3.64 (a) $r^2 = (0.596)^2 = 0.3552$. About 35.52% of the variation in the percent change for the entire year is accounted for by the linear model relating the percent change for the entire year to the percent change in January.
(b) When using the least-squares regression line with x = percent change in January to predict y = percent change for the entire year, we will typically be off by 8.3%.

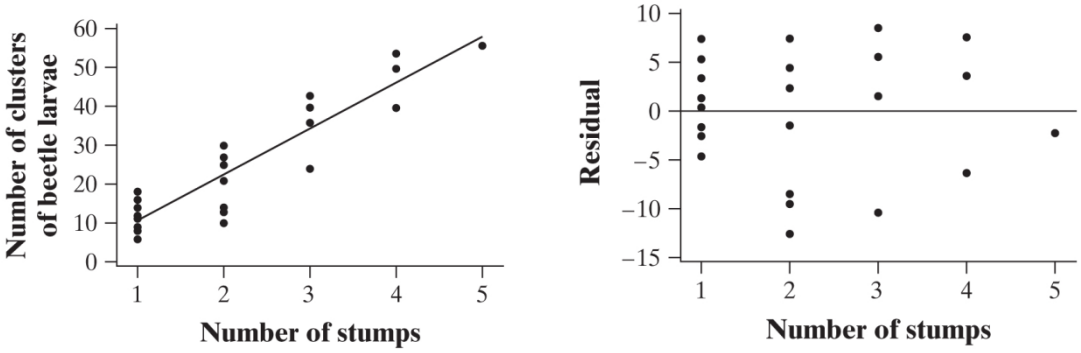
3.65 (a) $\hat{y} = x$, where y = grade on final and x = grade on midterm.
(b) A student with a score of 50 on the midterm is predicted to score $\hat{y}$ = 46.6 + 0.41(50) = 67.1 on the final. A student with a score of 100 on the midterm is predicted to score $\hat{y}$ = 46.6 + 0.41(100) = 87.6 on the final.
(c) These predictions illustrate regression to the mean because the student who did poorly on the midterm (50) is predicted to do better on the final (closer to the mean) while the student who did very well on the midterm (100) is predicted to do worse on the final (closer to the mean).

3.66 (a) $\hat{y} = x$

(b) A player with a first month batting average of 0.200 is predicted to have a rest-of-season batting average of $\hat{y} = 0.245 + 0.109(0.200) = 0.267$ A player with a first month batting average of 0.400 is predicted to have a rest-of-season batting average of $\hat{y} = 0.245 + 0.109(0.400) = 0.289$.

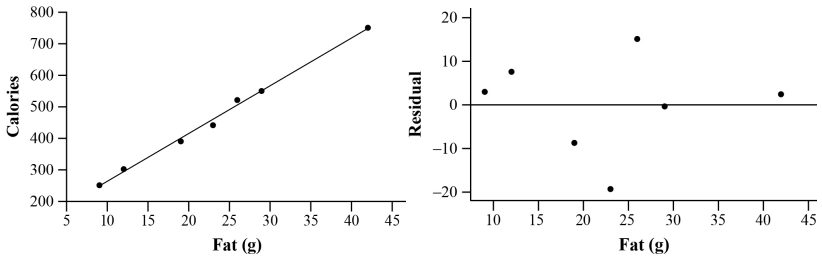
(c) These predictions illustrate regression to the mean because the player who hit poorly in the first month (0.200) is predicted to hit better the rest of the season (closer to the mean) while the player who hit very well in the first month (0.400) is predicted to do worse (closer to the mean).

3.67 *State*: Is a linear model appropriate for these data? If so, how well does the least-squares regression line fit the data? *Plan*: To determine if a linear model is appropriate, we will look at the scatterplot and residual plot to see if the association is linear or nonlinear. Then, if a linear model is appropriate, we will use the standard deviation of the residuals and r^2 to measure how well the least-squares line fits the data. *Do*: The scatterplot below shows a moderately strong, positive linear association between the number of stumps and the number of clusters of beetle larvae. The residual plot doesn't show any obvious leftover pattern, confirming that a linear model is appropriate.



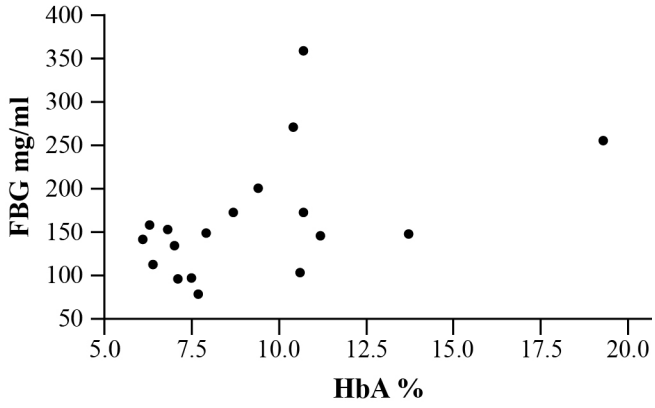
The equation of the least-squares regression line is $\hat{y} = -1.29 + 11.89x$, where y = number of clusters of beetle larvae and x = number of stumps. The standard deviation of the residuals is s = 6.42. This means that our predictions will typically be off by about 6.42 clusters when we use the least-squares regression line to predict the number of clusters of beetle larvae from the number of stumps. Finally, $r^2 = 0.839$, meaning 83.9% of the variation in the number of clusters of beetle larvae is accounted for by the linear model relating number of clusters of beetle larvae to the number of stumps. *Conclude*: The linear model relating number of clusters of beetle larvae to the number of stumps is appropriate for these data. Furthermore, the least-squares regression line fits the data well, accounting for more than 80% of the variation in number of clusters of beetle larvae.

3.68 *State*: Is a linear model appropriate for these data? If so, how well does the least-squares regression line fit the data? *Plan*: To determine if a linear model is appropriate, we will look at the scatterplot and residual plot to see if the association is linear or nonlinear. Then, if a linear model is appropriate, we will use the standard deviation of the residuals and r^2 to measure how well the least-squares line fits the data. *Do*: The scatterplot below shows a strong, positive linear association between fat and calories. The residual plot doesn't show any obvious leftover pattern, confirming that a linear model is appropriate.



The equation of the least-squares regression line is $\hat{y}$ = 110.44 + 15.1682x, where y = calories and x = fat. The standard deviation of the residuals is s = 12.25. This means that our predictions will typically be off by about 12.25 calories when we use the least-squares regression line to predict number of calories from grams of fat. Finally, r^2 = 0.996, meaning 99.6% of the variation in the number of calories is accounted for by the linear model relating number of calories to grams of fat. *Conclude*: The linear model relating number of calories to the grams of fat is appropriate for these data. Furthermore, the least-squares regression line fits the data very well, accounting for nearly all of the variation in number of calories.

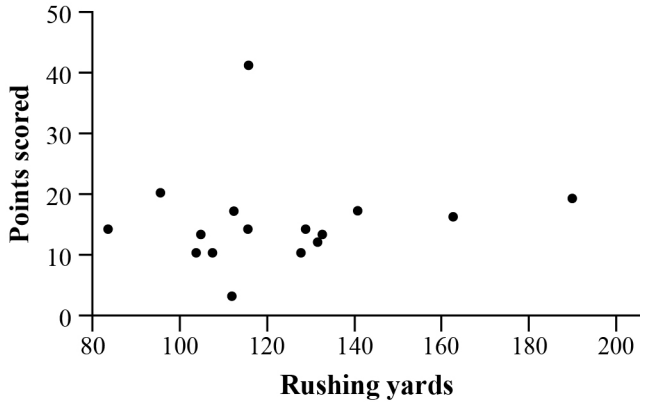
3.69 (a) A scatterplot of this relationship is shown below. There is a moderate, positive linear association between HbA and FBG. There are possible outliers to the far right (subject 18) and near the top of the plot (subject 15).



(b) Because the point for subject 18 is in the positive, linear pattern formed by most of the data values, it will make the correlation closer to 1. Also, because the point is likely to be below the least-squares regression line, it will "pull down" the line on the right side, making the slope closer to 0. Without the outlier, the correlation decreases from r = 0.4819 to r = 0.3837 as expected. Likewise, without the outlier, the equation of the line changes from $\hat{y} = 66.4 + 10.4x$ to $\hat{y} = 52.3 + 12.1x$.

(c) The point for subject 15 makes the correlation closer to 0 because it decreases the strength of what would otherwise be a moderately strong positive association. Because this point's x coordinate is very close to $\bar{x}$, it won't influence the slope very much. However, it will make the y intercept increase because its y coordinate is so large compared to the rest of the values. Without the outlier the correlation increases from $r = 0.4819$ to $r = 0.5684$, as expected. Likewise, without the outlier, the equation of the line changes from $\hat{y} = 66.4 + 10.4x$ to $\hat{y} = 69.5 + 8.92x$.

3.70 (a) A scatterplot of this relationship is shown below. There appears to be a very weak, positive association between points scored and rushing yards. With the exception of the outlier at 116 yards and 41 points, the association looks fairly linear.



(b) Because this point is in the positive, linear pattern formed by most of the data values, it will make the correlation closer to 1. Also, because the point is likely to be above the least-squares regression line, it will "pull up" the line on the right side, making the slope a little steeper. The correlation with the point is $r = 0.10$ and without the point the correlation drops to $r = 0.02$. Likewise, with the point the equation of the least-squares regression line is $\hat{y} = 11.4 + 0.031x$. Without the point the equation changes to $\hat{y} = 14 + 0.008x$.

(c) This outlier makes the correlation closer to 0 because it decreases the strength of what would otherwise be a moderately strong positive association. Because this point's x coordinate is very close to $\bar{x}$, it won't influence the slope very much. However, it will make the y intercept increase because its y coordinate is so large compared to the rest of the values. Without the outlier, the correlation increases from $r = 0.10$ to $r = 0.32$, the slope changes from 0.031 to 0.050, and the y intercept drops from 11.4 to 7.2.

3.71 a

3.72 a

3.73 c

3.74 a

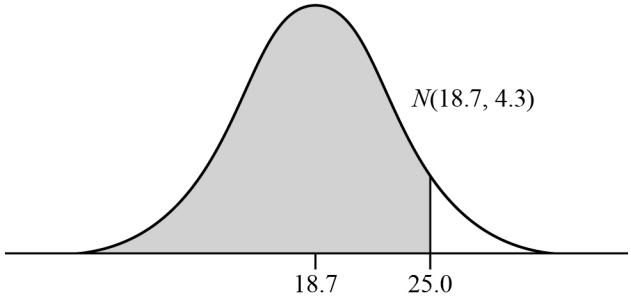
3.75 d

3.76 a

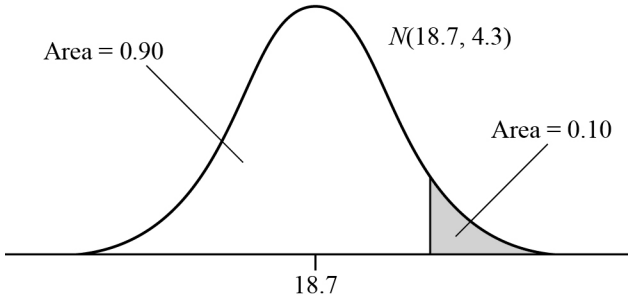
3.77 b

3.78 e

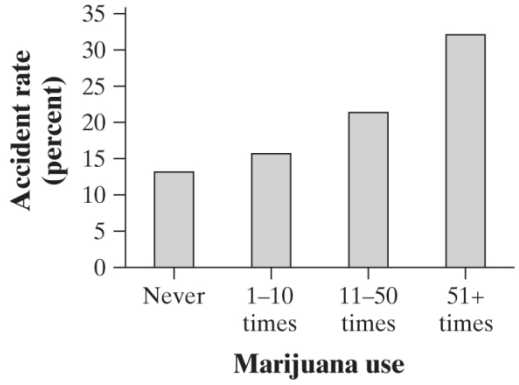
3.79 **Step 1: State the distribution and values of interest.** For these vehicles, the combined mileage follows a Normal distribution with mean 18.7 and standard deviation 4.3. We want to find the percent of cars with lower mileage than 25 (see graph below). **Step 2: Perform calculations. Show your work.** The standardized score for the boundary value is $z = \frac{25 - 18.7}{4.3} = 1.47$. From Table A, the proportion of z-scores below 1.47 is 0.9292. *Using technology:* The command normalcdf(lower: –1000, upper: 25, μ: 18.7, σ: 4.3) gives an area of 0.9286. **Step 3: Answer the question.** About 93% percent of vehicles get worse combined mileage than the Chevrolet Malibu.



3.80 **Step 1: State the distribution and values of interest.** For these vehicles, the combined mileage follows a Normal distribution with mean 18.7 and standard deviation 4.3. The boundary value that separates the top 10% from the rest of the distribution has an area of 0.90 to its left (see graph below). **Step 2: Perform calculations. Show your work.** Look in the body of Table A for the value closest to 0.90. A z-score of 1.28 gives the closest value (0.8997). Solving $1.28 = \frac{x - 18.7}{4.3}$ gives $x = 24.2$. *Using technology:* The command invNorm(area: 0.9, μ: 18.7, σ: 4.3) gives a value of 24.2. **Step 3: Answer the question.** The top 10% of all vehicles get at least 24.2 mpg.



3.81 (a) A bar graph is given below. There is evidence of an association between accident rate and marijuana use. Those people who use marijuana more are more likely to have caused accidents.



(b) Even if there is a strong association between two variables, we should not conclude that changes in one variable necessarily cause changes in the other variable. It could be that drivers who use marijuana more often are more willing to take risks than other drivers and that the willingness to take risks is what is causing the higher accident rate.